Hierarchical Clustering in Machine Learning

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as **hierarchical cluster analysis** or HCA.

In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the **dendrogram**.

A **dendrogram** is a diagram representing a tree. This diagrammatic representation is frequently used in different contexts:

- in hierarchical clustering, it illustrates the arrangement of the clusters produced by the corresponding analyses.
- in computational biology, it shows the clustering of genes or samples, sometimes in the margins of heatmaps.
- in phylogenetics, it displays the evolutionary relationships among various biological taxa. In this case, the dendrogram is also called a phylogenetic tree.

Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

The hierarchical clustering technique has two approaches:

- **Agglomerative**: This is a "bottom-up" approach: Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive**: This is a "top-down" approach: All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Agglomerative HC

STEP 1: Make each data point a single-point cluster ➞ That forms N clusters

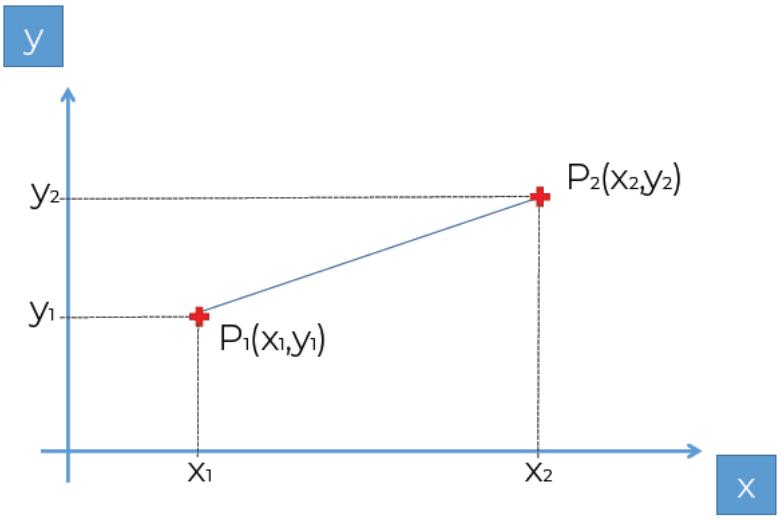
STEP 2: Take the two closest data points and make them one cluster ➞ That forms N-1 clusters

STEP 3: Take the two closest clusters and make them one cluster ➞ That forms N - 2 clusters

STEP 4: Repeat STEP 3 until there is only one cluster

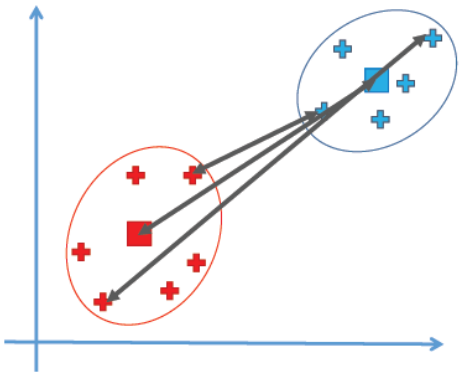
FIN

Euclidean Distance



Euclidean Distance between P_1 and $P_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Distance Between Clusters

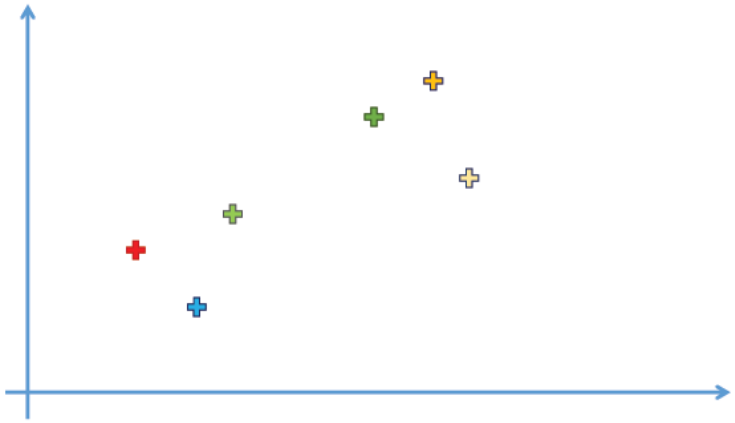


Distance Between Two Clusters:

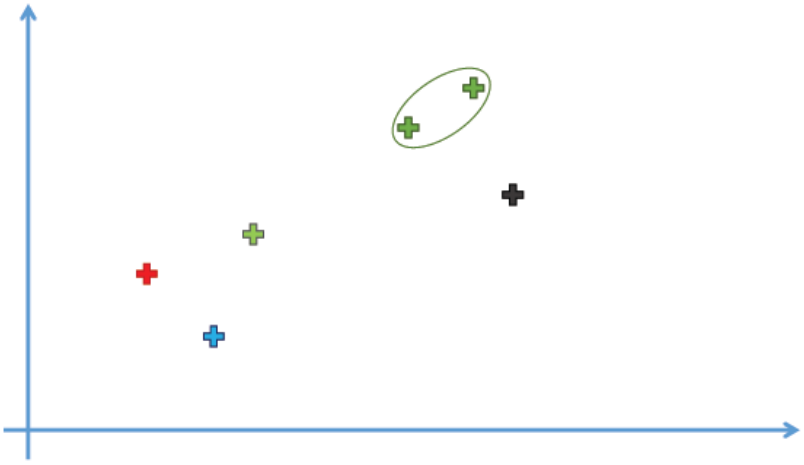
- Option 1: Closest Points
- Option 2: Furthest Points
- Option 3: Average Distance
- Option 4: Distance Between Centroids

Agglomerative HC:

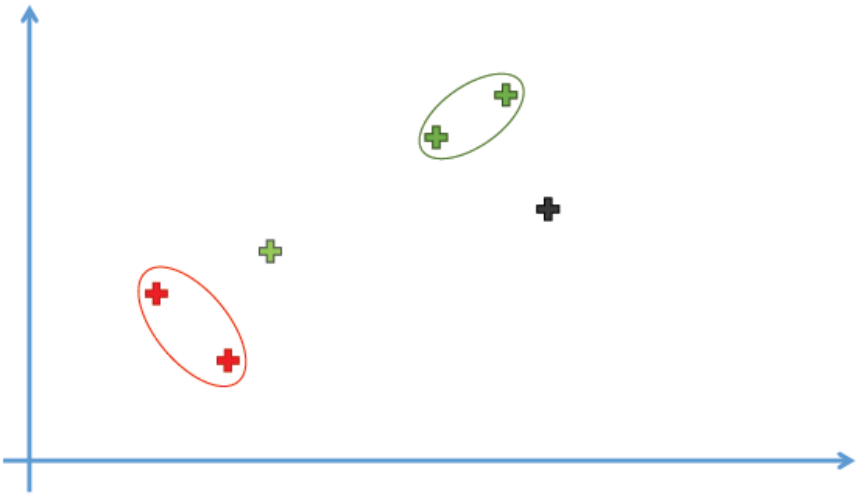
STEP 1: Make each data point a single-point cluster ⟶ That forms 6 clusters



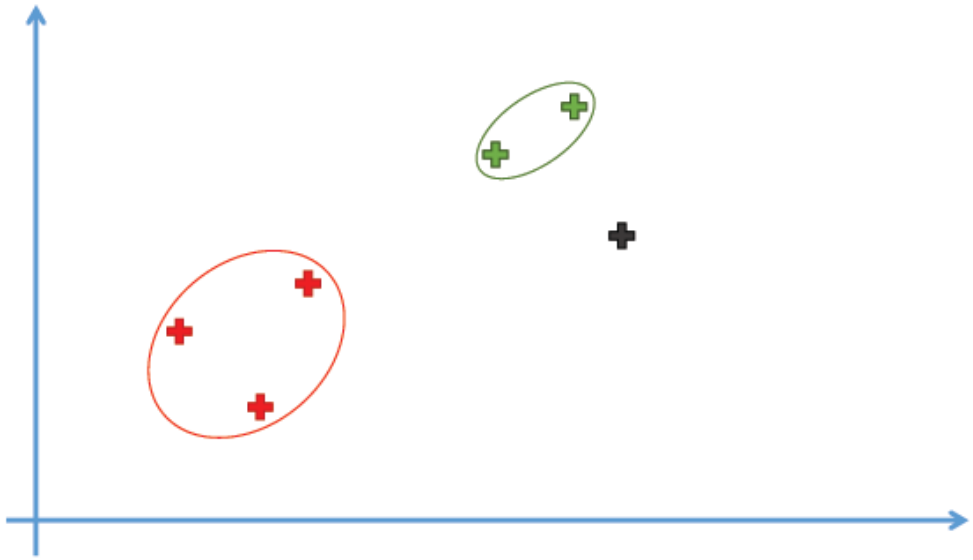
STEP 2: Take the two closest data points and make them one cluster
➞ That forms 5 clusters



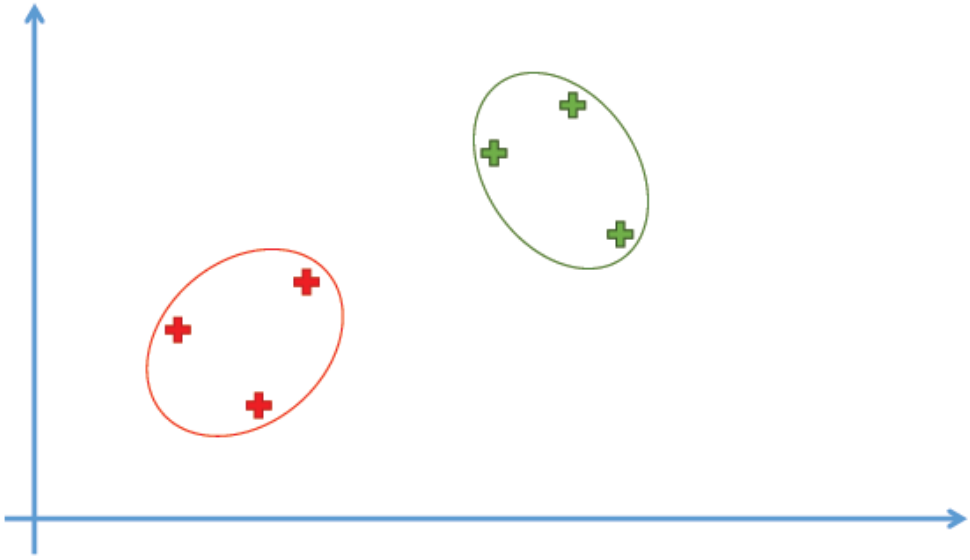
STEP 3: Take the two closest clusters and make them one cluster
➞ That forms 4 clusters



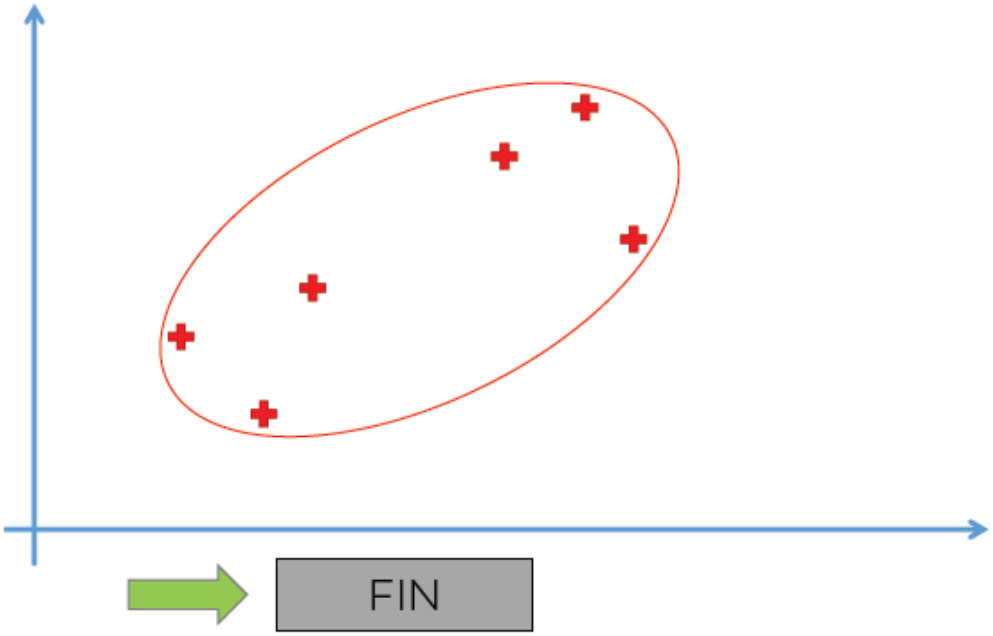
STEP 4: Repeat STEP 3 until there is only one cluster



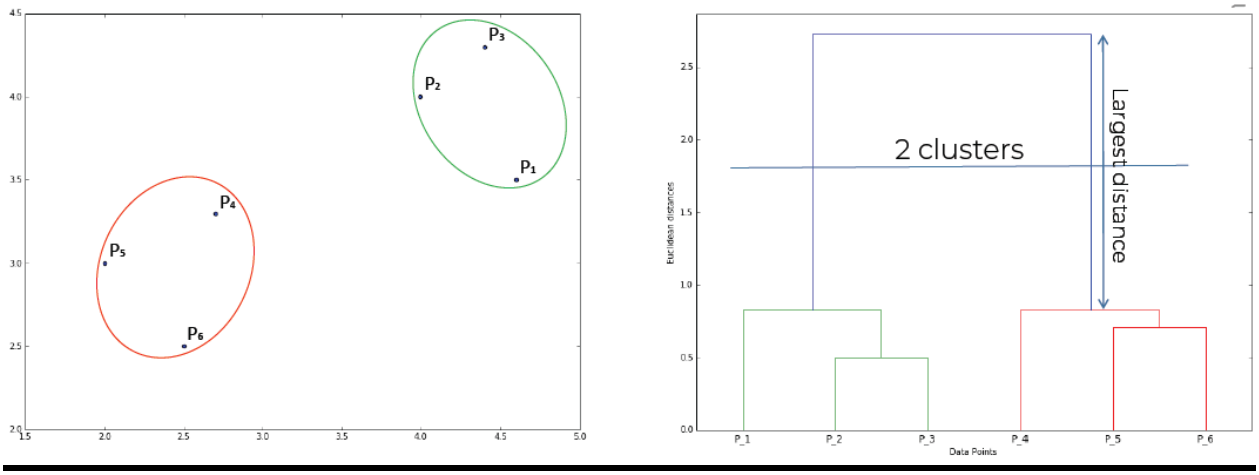
STEP 4: Repeat STEP 3 until there is only one cluster



STEP 4: Repeat STEP 3 until there is only one cluster



Dendrograms Optimal of the Clusters:



How do we use the dendrogram to get the most value out of the HC.

Look at Horizontal level and set distance threshold or height of the Euclidean distance. Set dissimilarity of the threshold mean each of the cluster dissimilarity is less than threshold. We can simply find the number of thresholds like how many vertical lines crossing the horizontal threshold that is the number of clusters. If the threshold level is very low then number of clusters may be six. So, how do we find the optimal number of clusters. What can tell the dendrogram might be a good guide for us to select the optimal clusters.

What are the standard approaches, highest vertical distance that we find the dendrogram mean any line that will not cross the horizontal line (extended). Green lines for P2 and P3 can be consider but P1 line not be consider because it crosses hypothetically horizontal green line. Same way P4, P3 and P6 not be considered because hypothetically same horizontal extended line crosses the all-vertical lines. Now find the longest violet color lines not crosses any horizontal lines. So, this is the largest distance therefore set a threshold that cross the largest distance. This approach telling the optimal cluster. Here optimal cluster is two clusters.